① Data manipulation at scale - Washington

week 1

All science is becoming data science, which has now lost of power. Anthropology, culture... are becoming hard sciences by the virtue of data-driven methods.

Mike Driscoll's three sexy skills of data geeks

1) Statistics (traditional analysis)
2) Data munging (parsing, scraping and formatting data)
3) Visualization (graphs, tools, etc.)

Three types of tasks (data scientist)

1) Preparing to run a model

Gathering, cleaning, integrating, restructuring, transforming, loading, filtering, deleting, combining, merging, verifying extracting, shaping, massaging

2) Running the model
3) Communicating the results

- Data science is about producing data products (data-driven application for example) that you may use or that other people will use. spellchecker...

Data product types: ① Data-driven apps ② Interactive visualizations (google flu application)

② Data manipulation at scale - Washington Week 1

③ Online databases (Enterprise data warehouse, Sloan Sky Survey)

- Data science is about building data products, not just answering questions.
- Data products empower others to use the data. They may help you to communicate your results or empower others to do their own analysis.

Distinguishing data science from Business Intelligence

Business intelligence systems: data warehouse and dashboards that use data from the data warehouse to answer specific questions.
Both of them require lots of effort to design and build. They are not too adaptable when requirements change.
A software stack built for BI may not be appropriate for data science where changes in requirements are considered the norm.
BI is associated with a particular approach to a particular set of problems and data science it in some sense broader. BI engineers are not expected to consume their own data products and perform their own analysis. Usually they are building tools for others to use. As a data scientist you do both: build data products and perform make decisions with the data analysis.

③ data manipulation at scale - Wrathnyka

Week 1

- Data scientists are less concerned with building permanent infrastructure for others to use, and more on answering questions and communicating the results
- Data science is software-intensive and data scientists create data products, but core infrastructure is usually developed by software engineering position.
- Data science is not limited by scale. Indeed, data scientists who only have experience working with "desktop-scale" data are not really competitive as data scientists.
- Business Intelligence is tied closely to the design, implementation, and use of data warehouses, and the database-oriented technology to support that process.
- Data scientists may make use of data warehouses but may also work with variety of less-structured sources.

Distinguishing data science with statistics

- Some of the methods/models from statistics are used by data scientists.
- However, statisticians assume that all data will fit in the main memory on a single machine. The field was born out of the need to extract the most information possible from a very sparse/expensive to collect and therefore very small dataset. That is not the problem anymore. There is a shift from a data poor regime to a data rich regime. The set of challenges therefore moves from a need of mathematics to squeeze information out of a dataset to

(4) Data manipulation at scale - Wathinpka week 1

engineer a process to handle very large datasets.

Distinguishing data science from Data (base) management.

Database experts / programmers / admin bring lots of skills appropriate for data science tasks but there is a focus on a particular data model which is usually the relational data model (rows and columns). Relational databases may not be the appropriate tools for many data coming from sources such as videos, audio, graphs...

Distinguishing data science from Visualization

Visualization experts bring lots of skills to the table. But they are less concerned with max and scale of the data (need many hundreds of machines).

Distinguishing data science from Machine learning

It is the closest to data science. The time that the data scientist spends on selecting a ML model and running it is very small compared to other tasks = data preparation.

"I worry that the Data Scientist role is like the mythical "webmaster" of the 90s: master of all trades."

Aaron Kimball, CTO Wibidata

⑤ Data manipulation at scale - Washington week 1

From related fields to data science

- If you are a DBA (DataBase administrator) you need to learn to deal with unstructured data.
- If you are a statistician, you need to learn to deal with data that does not fit in memory
- If you are a software engineer, you need to learn statistical modeling and how to communicate results.
- If you are a business analyst, you need to learn about algorithms and tradeoffs at scale.

Four dimensions of data science (cause)

① Breadth

tools ← ——— → abstractions (focus)

Hadoop
PostgreSQL
glm(..) in R
Tableau

MapReduce
Relational Algebra
Logistic Regression
InfoVis

② Depth

few structures ← ——— → statistics (manipulation of data)

Management
Relational Algebra

Analysis
Linear algebra

(6) Data manipulation at scale - Washington week 1

standards — ad hoc files

(3) Scale

Desktop — Cloud (hundreds of machines for ex.) — Tools

Main memory
R
local files

Distributed
Hadoop
S3, Azure storage

⚠ Scale is very important to be equipped well as a data scientist.

(4) Target

Hackers — Analysts — Tools

Assume proficiency in Python, Java, R

Assume little or no programming

Tools vs. abstractions

Databases and what is going on in the NoSQL systems

Pre-2004: Commercial RDBMS (Big3) + open source (MySQL)

2004: Paper from Dean et al. on Map Reduce (Google)
It allowed to process very large datasets and reboot the features set.

⑦ Data manipulation at scale - Washington Week 1

MapReduce really focused on just scale out parallelism and none of the other database features. This was exciting for some people because they didn't have to use the extra features they didn't need nor pay the extra fee.

2008: implementation of this paper (MapReduce) with the Open source Hadoop 0.17 release (people from Yahoo)

2008: Olston et al. Pig: Relational Algebra programming environment for Hadoop. It is the secret sauce within relational databases. Widely successful and needed: have relational style programming on top of this non-relational system.

2008: DryadLINQ: Relational Algebra in a Hadoop-like system. Competitor of Pig (similar).

2009: Thusoo et al. HIVE: SQL on Hadoop

2009: HBase: Indexing for Hadoop (what databases have)

2010: Dietrich et al. Schemas and Indexing for Hadoop

2012: Transactions in HBase (plus VoltDB, other New-SQL systems). Important topic.

MapReduce permanent contributions:

- Fault tolerance: during a query processing, you don't lose data if the system fails

(8) Data manipulation at scale - Washington week 1

Previous systems have been built to work on a single machine. Now that we can run a query on hundred of machines the odd that one fail during a long query is high.

- Schema-on-Read: Until you can fit your data to the schema you have previously defined (for the data) you cannot work with. Now, we have very large amount of data that cannot really fit on predefined schema. Having to load your data into your database before you can touch it was a non-starter for many applications.

User-defined functions that don't suck:

Relational database vs NoSQL systems

What are the abstractions of data science?

Matrices and linear algebra? (Maths)
Relations and relational algebra?
Objects and methods?
Files and scripts?
Data frames and functions?

most valid options

There might be other options.

Desktop scale vs Cloud scale

Large set of data is not just bigger, it is different. It requires a different way of thinking, techniques and stack of technology.

⑨ Data manipulation at scale - Washington week 1

Data science is not just about function, model manipulation of data technique but also about scale. Scale is a fundamental limitation of a whole category of technology (R included even if there is some advancement on that). You would be out of luck if your data doesn't fit in your memory. You could use indices to limit search or parallel data search and analysis.
It is good to be aware of what other systems can do and be able to take advantage of those flexibly.
Writing your applications in terms of those applications that do scale out is a critical skill in data science.

(a bit outdated but to give an idea):
Current practice based on data download (FTP/GREP)
Will not scale to the datasets of tomorrow.

- You can GREP 1M in a second
- You can GREP 1GB in a minute
- You can GREP 1TB in 2 day
- You can GREP 1PB in 3 years

} more and more costly
You have to search in a smarter way.

GREP: search a file linearly for a particular pattern.
Every single line of a file.

(Sparse) Matrix Multiply in SQL

Representation of the matrix:

row id	column id	value

(10) Data manipulation at scale - Washington week 1

```
SELECT A.row_number, B.column_number. SUM (A.value *
                                           B.value)
FROM A, B
WHERE A.column_number = B.row_number
GROUP BY A.row_number, B.column_number
```

Representing matrices inside a database may be inefficient if the matrix is dense but if the matrix is sparse it can be efficient

sands unusual.

A fourth paradigm of science

Empirical: you observe the natural world and you replicate it. (for thousands of years)

Theoretical: in the last few hundred of years, theoretical models have been accepted. You experiment these models.

Computational: in the last 50 years, high speed computation has allowed humans to solve problems very fast that weren't possible before

→ Three methods of inquiries (traditional) scientific

eScience (4th method of scientific inquiries) in the last 10 years. It requires very large datasets through new methods and infrastructures. It is about massive and complex data, it requires automated or semi-automated techniques. The tools here are the same for

(11) Data manipulation at scale - Washington week 1

data science.

Science is about asking questions.

Traditionally: "Query the world"
Data acquisition activities coupled to a specific hypothesis

eScience: "Download the world"
Data acquired en masse in support of many hypotheses
The data acquisition can be done independently to the hypotheses in some cases.

- The cost of data acquisition has dropped precipitously thanks to advances in technology
- The cost of finding, integrating and analyzing data, then communicating results, is the new bottleneck. (data science.)
- The rate at which data is now produced has way outpaced the rate at which we can analyze it and ask Q.

Data intensive science example: The Web

20+ billion web pages x 20 kB = 400+ TB
One computer can read 30-35 MB/sec from one disk
= 4 months just to read the web.

eScience is about the analysis of data

- The automated or semi-automated extraction of knowledge from massive volumes of data
- The main tools for looking for answers are algorithms and technology

(12) Data manipulation at scale - Washington week 1

- There is simply too much of it to look at
- But it's not just a matter of volume
- The three V's of Big Data
  - Volume: number of rows / objects / bytes
  - Variety: number of columns / dimensions / source
  - Velocity: number of rows / bytes per unit time
- More V's:
  - Veracity: Can we trust this data?

Summary - eScience

- Science is in the midst of a generational shift from a data poor enterprise to a data-rich enterprise
- Data analysis has replaced data acquisition as the new bottleneck to discovery
- What does this have to do with Business?

Business is beginning to look a lot like science

- Acquire data aggressively and keep it around (in case of usefulness)
- Hire data scientists
- Make empirical decisions

You can take the lessons learned in science and apply them to business and vice-versa: one thing science is behind business is in adopting new technology (there has been proportionally less spend in IT in science than in business)

(13) Data manipulation at scale - Wednesday week 7

Big data: Three challenges

- **Volume**: size of the data (in bytes, rows, objects.. vertical dimension of the data)
- **Velocity**: the latency of data processing relative to the growing demand for interactivity.
How fast data is coming compared to how fast it needs to be consumed?
For lots of application, interaction time is increasingly important.
- **Variety**: diversity of sources, formats, quality, structures. Integration of all of these sources is very important and can require lots of your time (data cleaning & co = 90% of data scientists sometimes)

All of three challenges relevant in deep data science

Big Data

"Big Data is any data that is expensive to manage and hard to extract value from" Michael Franklin, University of Berkeley.

- Idea: "Big" is relative. There is no cutoff. "Difficult Data" is perhaps more appropriate. It is about being challenging

(14) Data manipulation at scale - Wattenhofer week 1

- Private data is becoming commodified

Emerging topics in data science: ethics, validation, privacy

"... the necessity of grappling with Big Data, and the desirability of unlocking the information hidden within it, is now a key theme in all the sciences - arguably the key scientific theme of our times."

Francis X. Diebold, University of Pennsylvania

Assignment 1: Twitter sentiment analysis

Unicode: standard for representing a much larger variety of characters beyond the roman alphabet (greek, russian, mathematical symbols, logograms from non-phonetic writing systems such as kanji...)

Unicode string: u "This is a string" (prefixed with the letter "u").

Encode method if error involving printing unicode:

```
unicode_string= u"aaàççñññ"
encoded_string = unicode_string.encode('utf-8')
print encoded_string
```